The Knowledge Engineering Review, Vol. 19:2, 133–146. © 2004, Cambridge University Press
doi:10.1017/S0269888904000190 Printed in the United Kingdom

A review of explanation methods for heuristic expert systems

CARMEN LACAVE[1] and FRANCISCO J. DIEZ[2]

[1]*Department of Computer Science, University of Castilla-La Mancha, Paseo de la Universidad, 4, 13071 Ciudad Real, Spain; e-mail: carmen.lacave@uclm.es*
[2]*Department of Artificial Intelligence, UNED, Avda. Juan del Rosal, 16, 28040 Madrid, Spain; e-mail: fjdiez@dia.uned.es*

Abstract

Explanation of reasoning is one of the most important abilities an expert system should provide in order to be widely accepted. In fact, since MYCIN, many expert systems have tried to include some explanation capability. This paper reviews the methods developed to date for explanation in heuristic expert systems.

1 Introduction

It is well known that explanation constitutes a very important topic to be addressed by any software product. In the case of expert systems, explanation is even more important because without it users would not be able to reject a system's recommendation when it is wrong and they would be reluctant to accept its advice even when it is right (Wooley, 2000; Doyle *et al*., 2003; McGuinness & Pinheiro da Silva, 2003). Therefore, human–computer collaboration requires mutual understanding: machine models must take into account human cognitive processes and at the same time make artificial reasoning understandable to human users (Buchanan & Shortliffe, 1984; Henrion & Druzdzel, 1990; Ye & Johnson, 1995). During the deployment of an expert system, the explanation capability is very useful for convincing the user of the correctness of the results; in particular, physicians are very reluctant to accept the advice of a machine if they cannot understand how it was obtained (Teach & Shortliffe, 1984). In the case of tutoring systems, the explanation capability is essential for improving the student's skills.

It is surprising, however, that the amount of research devoted to this subject has been relatively small compared with other areas of artificial intelligence. There are only isolated pieces of work, seldom used in real-world applications; in fact, the explanation capability of most of today's expert systems and packages is even poorer than that of MYCIN, which is often regarded as the first expert system.

With the purpose of including an explanation method for Prostanet (Lacave & Díez, 2003), a causal Bayesian network for diagnosing prostate cancer, we decided to analyse the explanation methods that have been developed, not only in Bayesian networks but also in heuristic expert systems. As a result of such work, we reviewed the explanation methods for Bayesian networks (Lacave & Díez, 2002). Therefore, the aim of this paper is to analyse what has been done to date and what remains to be done in the field of explanation in heuristic expert systems. However, we will not consider the field of Case Based Reasoning systems because there is a complete review in Doyle *et al*. (2003). Furthermore, a thorough study of the evolution of explanation systems, not only in computer science but also in other fields such as linguistics, logicians, philosophers, etc. can be found in Moulin *et al*. (2002).

After introducing very briefly the fundamental properties of explanation (Section 2), we review the methods of explanation for the particular field of heuristic expert systems (Section 3) and conclude by pointing out possible lines for future research (Section 4).

2 Concept and properties of explanation

It has been demonstrated (Buchanan & Shortliffe, 1984; Henrion & Druzdzel, 1990; Ye & Johnson, 1995; Moulin *et al*., 2002) that explanation plays a very important role in the acceptance of expert systems. However, one of the most important problems arises when trying to define an explanation or when specifying which properties it should have to be 'good enough', because explanations which are not properly tailored for different kinds of users may be worse than providing no explanation at all (Cawsey, 1993). In this section we summarize the conclusions of a previous work (Lacave & Díez, 2002) by presenting a definition for the concept of explanation as well as some properties that characterize explanation methods.

2.1 Definition

From the origins of philosophy and science, many researchers of both disciplines have tried to determine the meaning of explanation, connecting it with comprehension and description. Also in the field of expert systems there are different interpretations of its meaning. For instance, the rule-based expert system MYCIN (Shortliffe *et al*., 1973) is able to show *how* it obtains a conclusion or *why* it is requesting additional information from the user. In probabilistic expert systems the 'best explanation' is defined as the most probable assignment of values to a set of variables (Pearl, 1988). Other systems try to offer a simple but comprehensible report about the domain and the reasoning (Swartout, 1983; Reggia & Perricone, 1985; Horvitz *et al*., 1986; Strat, 1987; Langlotz *et al*., 1988; Chandrasekaran *et al*., 1989; Wick & Thompson, 1992; Carolis *et al*., 1996). Finally, in the most sophisticated methods, explanation constitutes an 'intelligent' dialogue with the system user through natural language by way of interactive methods (Carenini *et al*., 1994; Moore, 1994; Cawsey, 1995; Fiedler, 2001a).

After analysing the literature on explanation in artificial intelligence (Wick, 1989), we concluded that *explaining* consists in *exposing something* in such a way that it is *understandable* for the receiver of the explanation—so that he/she improves his/her knowledge about the object of the explanation—and *satisfactory* in that it meets the receiver's expectations (Lacave & Díez, 2002).

2.2 Properties

Another consequence of such a study was the identification of several properties of explanation methods (Lacave & Díez, 2002)—characteristics (summarized in Table 1) that correspond to the main concepts on which our definition is based: content, communication and adaptation.

2.2.1 Content

One of the most important aspects of explanation is related to *what* is going to be explained, which depends on the following:

Focus of the explanation In any expert system, there are three basic issues to be explained:

- the available evidence, if any;
- the knowledge base; and
- the reasoning process performed by the system to obtain (or not) a conclusion.

Table 1 Main properties of explanation methods

Category	Property	Options
Content	Focus	evidence/model/reasoning
	Purpose	description/comprehension
	Level	micro/macro
	Causality	causal/non-causal
Communication	User–system interaction	menu/predefined questions/natural language dialogue
	Display of explanations	text/graphics/multimedia
Adaptation	User's knowledge about the domain	no model/scale/dynamic model
	User's knowledge about the reasoning method	no model/scale/dynamic model
	Level of detail	fixed/threshold/auto

The explanation of *evidence* consists in determining which hypothesis justifies the observed findings. For instance, in medical expert systems, such an explanation could consist in determining the disease or diseases that justify the symptoms, signs, test results, etc. of a certain patient.[1]

The explanation of the *model* consists in displaying (verbally or graphically) the information contained in the knowledge base with the aim of assisting human experts in the *construction* and *debugging* of the expert system or of offering a user some knowledge about the domain for *instructional purposes*.

The explanation of *reasoning* may provide justification for the *results obtained* by the system and the *reasoning process* that produced them. During the evaluation of the expert system, the explanation of the reasoning process is an invaluable tool for isolating and correcting inaccurate pieces of information in the knowledge base. On the other hand, it may also be of interest, especially for educational systems, to explain the *results not obtained* by the system; in particular, the findings that oppose such a conclusion and the findings that would be necessary to support it. Finally, any explanation tool should also include the ability to perform *hypothetical reasoning*, in order to ascertain what results the system would have returned if one or more of the observed variables had taken on different values.

Purpose There are two different goals an explanation capability may try to accomplish, leading to two types of explanation:

Description. This kind of explanation consists in showing the underlying *knowledge base* (in the case of model explanation), or providing *further details* on the *conclusions*, or displaying *intermediate results* (in the case of reasoning explanation).

Comprehension. In this case, the explanation tries to make the user understand the implications of the model or the conclusions of the system and/or the *relation* between them. For instance, how each finding affects the conclusion, either individually or in conjunction with other findings.

Level Sember & Zukerman (1990) proposed another classification of explanation methods for Bayesian networks, which can also be applied to other types of expert system:

Micro level. A micro-level explanation consists of a detailed justification of the variations produced in a particular variable (or hypothesis) as a consequence of the variations in its

[1] However, in heuristic expert systems, contrary to the case of Bayesian networks, this 'explanation of evidence' is not considered as an explanation facility, since the generation of a hypothesis is a core activity of these systems.

neighbours. In the case of a rule-based expert system, the micro level would consist in analysing the variables contained in a certain rule, or the rules containing a certain variable.

Macro level. This level of explanation analyses the main lines of reasoning (the paths in the Bayesian network, the chains of rules, etc.) that lead from the evidence to a certain conclusion. A macro-level explanation for a traditional expert system would consist of one or several chains of rules.

Causality There are several reasons for using causal models in artificial intelligence. First, human beings tend to interpret events in terms of cause–effect relations (Kahneman *et al.*, 1982; Pennington & Hastie, 1988). Therefore, causal models are easier to construct and modify (Pennington & Hastie, 1988; Henrion, 1989), and also more easily understood by users (Suermondt, 1992; Druzdzel, 1993). Second, the identification of *invariant*[2] causal relationships in a domain allows the prediction of effects of both spontaneous causes (usually corresponding to random variables) and actions (sometimes called manipulations or interventions) (Pearl, 2001). Third, the concept of explanation is very closely tied to the notion of causation; in fact, one of the modalities of scientific explanation consists of finding the causes of the observed facts. Therefore, some explanation methods for expert systems are specifically designed for causal models, while other methods are general, in the sense that they do not assume a causal interpretation of the model.

2.2.2 Communication

This is related to *how* the explanation is going to be offered to the user. This involves the interaction between the user and the system and the way it is going to be presented. In some systems, users can request explanations by selecting some variables or options from a certain menu. In other systems, users can pose questions and there are also some systems which try to offer a natural language dialogue. Alternatively, some systems only allow the user to ask for an explanation after the program has presented its conclusions, while other systems allow the user to interrupt the process of inference and request an explanation. The display of explanation can be made verbally, graphically, or in a hypertext and multimedia environment that combines interactive explanations with images, video and sound.

2.2.3 Adaptation

It is very important to consider to *whom* the explanation is offered in order to cover his/her expectations. With this aim, the system should take into account the *knowledge* the user has about the modelled *domain* and the *reasoning method* performed by the system. Most explanation methods do not take into account the variability of knowledge between different users and explanations are generated for a hypothetical user having a certain knowledge: some of them are intended for beginners, others assume the user is an expert.

Other methods rely on a *static user model*, i.e. they do not consider that the user's knowledge increases as he/she interacts with the system. In many cases, the user model often reduces to using a *scale* with two or three categories ('novice, advanced, expert'). In theory, an expert system should have the capability of classifying users automatically, but in practice, the user generally has to classify himself/herself on this scale at the beginning of his/her interaction with the system.

There are also a few methods which possess a *dynamic model* for each user, which allows them to adapt the explanation to the knowledge the user has at each moment.

Very close to this question is the *level of detail* of the explanation, which should be automatically adapted during the performance of the system. Unfortunately, only a few methods include this ability. Most of them consider a fixed level of detail and in some cases they offer a rudimentary possibility for modifying the threshold so that the explanation module only displays those items above it.

[2] 'Invariant' means that when one mechanism is subjected to changes, the others remain intact.

3 Explanation methods for heuristic expert systems

In this section we review the techniques developed over the last years for generating explanation in heuristic expert systems, in the light of the properties discussed above. One of the main properties required for every kind of 'intelligent program' is adaptation. Therefore, we divide the methods into two groups, according to their ability to *adapt* their behaviour to different user requirements.

3.1 Non-adaptive methods

The main characteristic of these methods is that the system does not have the ability to update its knowledge about the user during its performance. This implies that generally the *level of detail* of explanations is *fixed*. In some of the most recent systems (see below), the user has the ability to select the level of detail among a predefined small set of values (at most five). Although most of them do not consider any *user model*, some recent expert systems have the option of setting the user's level of knowledge about the modelled domain before their performance. However, the *interaction* between the system and the user is made through *predefined questions or options*. We present the most relevant works in chronological order.

3.1.1 Canned text

One of the first expert systems offering a rudimentary explanation tool was the one developed by Bleich (1972) to evaluate electrolyte and acid-based disorders. The technique for generating explanations was the one known as 'canned text', consisting of coding the answers to the queries the user could raise on the program source. Then, given certain evidence, the system would show the associated answers and the values of some variables. The main problem with this method is that the knowledge engineer has to foresee all the possible questions, which is really complicated for real world systems. As we can see, the purpose of this kind of method is the *comprehension* of the *reasoning* at a *macro level* and the explanations are presented *textually*.

3.1.2 MYCIN

MYCIN (Shortliffe *et al.*, 1973), an expert system for the diagnosis and therapy recommendation of infectious diseases, is considered to be the first expert system having a real explanation capability. It is a rule-based system whose question–answering capability offers three types of explanation: the reasoning performed by the system at a given time and both the static and the dynamic knowledge bases (Scott *et al.*, 1984). Moreover, it is also capable of justifying the selected therapy, including an explanation on the decided dose (Bennet & Scott, 1984). The main idea was to translate into 'natural language' the rules used during the reasoning process. To do this, each time the system poses a question, the user can ask for two kinds of explanation:

- *Why* [is the system requesting this information?]. In this case, the system offers a translation into natural language of the goals reached until that moment and those it is trying to find out, as well as the rules used in the process. For instance, if MYCIN asks 'What is the morphology of ORGANISM-1?' and the user types 'WHY?', MYCIN would rephrase this question as 'Why is it important to determine the category of ORGANISM-1?' and then explain the objective 'This will aid in determining the morphology of ORGANISM-1' showing a rule having the morphology as a premise and the category as a conclusion.
- *How* [has the system arrived at this conclusion?] Similarly, the system provides a translation into natural language of the specific rule. For example, if MYCIN says 'The category of ORGANISM-1 is pneumococcus' and the user types 'HOW?', MYCIN would rephrase this question as 'How was it established that ORGANISM-1 is pneumococcus?' and then show the rules that led to the conclusion.

Therefore, the main purpose of this method is the *comprehension* of both the *model* and the *reasoning*. The explanations are offered at the *micro level*, because at each moment the focus is made on one specific rule. The presentation of explanations is *textual*, using *natural language* and expressing the different degrees of certainty also by *numbers*.

However, the explanation capability of MYCIN has two important shortcomings. First, the system is not sure whether users understand the process of rule chaining because sometimes different sources of knowledge are mixed, for example, static and dynamic. Second it is very difficult to get a coherent argument when the number of chained rules is high.

3.1.3 Causal knowledge

In order to solve these deficiencies, new expert systems were developed in the Stanford Heuristic Programming Project. One common property in some of them is that they include *causal* knowledge. We describe them, according to their relevance.

NEOMYCIN This expert system was developed with the main purpose of explaining the results provided by MYCIN (Clancey, 1983, 1993). This system includes explicit causal knowledge, represented by meta-rules and a taxonomy of diseases. The goal is to explain not only the chain of reasoning but also the *reasoning strategy*. Therefore, the purpose is the *comprehension* of both the *model* and the *reasoning*. Explanations are generated at the *micro* and *macro* level and presented *textually*, in natural language.

Causal MYCIN Another improvement to MYCIN is the method for generating causal explanations developed by Wallis & Shortliffe (1984). They represented the knowledge base as a net of rules, which serves both to acquire knowledge and to generate tailored explanations, because it takes into account the user model and the level of detail required by him/her. With such a purpose, they associated a *complexity measure* and an *importance factor* to the rules and to the concepts related to the system conclusions. Moreover, they defined two parameters: the *user experience*, to represent his/her level of knowledge, and the *level of detail* of the explanations. By combining these parameters, the rules and the concepts to be included in the explanation can be selected dynamically depending on the different user needs. The purpose of this method is the *comprehension* of both the *model* and the *reasoning*. Explanations can be generated both at *micro* and *macro* levels and are presented *textually*, by natural language expressions.

XPLAIN In a different way, Swartout (1983) designed a methodology for developing his expert system XPLAIN to improve the explanations provided by the Digitalis Therapy Advisor (Silverman, 1975). The main idea was based on using an automatic programmer for both creating the expert system and explaining its reasoning by refinement from abstract goals. XPLAIN has the following elements:

- The domain model, which contains the descriptive facts of the domain and is represented by a causal net.
- The domain principles for reasoning, defined by several methods and heuristics.
- The automatic programmer for generating the expert system with the knowledge provided by the two kinds of domains.
- The structure needed for generating explanations, represented by a *tree*, which was developed during the construction of the expert system. Since it is generated by a refinement process from general to specific objectives, each node of the tree represents a refinement of the objective represented by its parent. The leaves of the tree define the basic operations the program has to produce.
- The translator for expressing in natural language the explanations, which takes the previous tree as its input. It has two elements: a *phrase generator*, for building phrases directly from the knowledge base; and an *answer generator*, which determines what parts of the structure should

be included in the explanation and the level of detail, taking into account the user needs and the state of the system.

Then, the objective of this method is the *comprehension* of the *model* and the *reasoning* by explanations generated at the *macro* level, which are presented in *natural language*. Moreover, the *level of detail* is modified during the performance of the system, depending on the user needs.

ABEL This system, developed by Patil *et al.* (1982) for managing acid–base and electrolite disorders, is based on the ideas of XPLAIN (described above), but the main difference is that several kinds of relations are distinguished: *cause–effect*, *associated–to* and *groupation*. The main objective is the *comprehension* of the *model* and the *reasoning* by explanations generated at the *macro* level, which are presented in *natural language* with five different *levels of detail*.

3.1.4 Reconstructive explanation

On a different level, we can find the proposals of Wick *et al.* (Wick & Slagle, 1989; Wick & Thompson, 1992) about *reconstructive explanation*, through which the process of reasoning and explanation are considered as two different and independent tasks. The main argument used is that when humans give explanations about how a problem has been solved, they do not describe every step followed to reach the solution. Moreover, sometimes other arguments which did not take part in the process can appear in order to improve the explanation. Therefore, they consider that the methods for generating explanations should take as their input the reasoning process performed by the system to get a solution and produce as output its explanation, taking into account an *explanation domain*.

These ideas have been put into practice in the REX prototype (Wick & Thompson, 1992) designed to explain how an expert system gets its results. The idea is as follows: the expert system produces a set of clues about the reasoning process which are filtered to determine the parts of the reasoning to be passed to the module for generating the explanation, taking into account the problem restrictions. This module uses the explanation domain to obtain the concepts of the explanation which justify the reasoning clues and then another module translates them into natural language. Therefore, the main objective is the *comprehension* of the *reasoning* by explanations generated at the *macro* level, which are presented in *natural language*.

3.1.5 Text planning

The problem of presenting explanations constitutes a topic of interest so that users' expectations are met, although it is not considered to be the main aim when providing explanations. In fact, the explanations presented in natural language are not very satisfactory, in the sense that in most cases the text produced has no cohesion or coherence. However, nowadays the generation of natural language expressions constitutes in itself a very important topic of research in artificial intelligence, following the development of some interesting proposals like *text planning* (Moore, 1994; Cawsey, 1995) for providing expert systems with explanation capabilities. Although XPLAIN can be considered as one of the first approaches in this field, we also describe the following systems because of their improvements relating to the natural language generation in the field of medicine. In both cases the main objective is the *comprehension* of the *model* and the explanations are presented at the *macro level* in *natural language*.

HEALTHDOC This system was developed by Marco *et al.* (1995) with the main purpose of generating medical reports for patients and materials for health education, *adapted* to the specific needs of each one. The system has the following properties:

- the reports and materials are presented textually; therefore there is *no interaction* between the user and the system;

- the user can print the documents or edit them in order to carry out some modifications;
- to adapt the text to different patients, the system has a database containing information about every patient, that is, their clinical data and their personal and cultural characteristics;
- it is possible to generate the text in *several languages*.

Each report is generated from a 'master document' that contains all the information to be presented in the report, including illustrations, annotations, etc., considered relevant by the doctor. It is represented in SPL (Sentence Plan Language) and is used as an intermediary for generating natural language. Moreover, each phrase is marked with coherence relations and references to other phrases. The master document is edited by the doctor with a tool provided by the system and then some of the content is selected to determine a plan of the text which will generate the natural language discourse, taking into account the rhetorical relations needed to produce it. During the whole process, the patient data are taken into account in order to adapt the report to his/her specific characteristics.

OPADE In a similar way, Carolis *et al.* (1996) have developed this expert system for generating recipient-centred explanations about drug prescription in the same way the doctor would do, taking into account the user requirements. Two kinds of user are considered: *direct* users, who directly interact with the system (doctors or nurses, in the case of medicine), and *indirect* users, who only receive a report of the results (patients). The system has four elements:

- the user model, containing the main characteristics of both kinds of user, direct and indirect;
- a text planner, consisting of several planning operators whose main objective is to build a tree containing the discourse plan based on the objectives and sub-objectives to be reached, depending on the user model;
- a module for solving conflicts between the two kinds of user;
- a text generator, based on augmented transition nets, which takes as input the tree generated by the text planner and produces a discourse formed by several phrases expressed in natural language.

3.1.6 Communication domain

Related to the ideas of text planning and reconstructive explanation, there are other proposals (Barzilay *et al.*, 1998) for generating expert-system explanations which are based on the classification of types of knowledge into *reasoning knowledge*, *communication domain knowledge* (i.e. knowledge about the domain), and, distinguishing it from the rest, *communication domain knowledge* (i.e. knowledge about how to communicate in the domain). This kind of knowledge acts as a bridge between the other two, facilitating both the development of the expert system and the text planner. The objective of this method is to help in the *comprehension* of the *model* and the *reasoning* by *natural language* explanations at the *macro* level.

3.2 Automatic adaptive methods

One of the key factors for an explanation to be satisfactory is that the persons involved in the process (the explainer and the explainee) are able to interact dynamically. This means that the receiver can ask the explainer whenever he/she needs to cover his/her expectations and, reciprocally, the explainer can update his/her/its discourse depending on the level of knowledge acquired during the process by the receiver. To some extent, the question of the communication from the user to the explainer has been addressed in certain systems by supplying different levels of detail or letting the user ask the system again once the explanation has been offered. However, none of the systems described supports the possibility of automatically adapting its level of knowledge about the user during its performance. In order to solve this deficiency, some new methods have been proposed, which we will describe in this section.

3.2.1 Reactive explanations

One of the most important researchers in this area is J. Moore (Moore & Swartout, 1990; Moore, 1994). She considers an explanation as a type of conversation between the user and the system and, therefore, she proposes a technique of *reactive explanation*, based on the ideas developed by Swartout (Section 3.1). This implies that the system has to know what has been said until each moment and to be able to react to any question made by the user in a clever way, identifying his/her objectives. In brief, the process is as follows: when the user poses a question or introduces a requested piece of information, an objective is generated, representing the abstraction of the topic to be included in the explanation. This objective is passed to the text planner, which identifies the explanation strategies that verify that objective. Then, the strategy that best fits the objective is found by heuristic techniques, taking into account the user model, the dialog until that moment, etc. Therefore, a tree is generated representing the plan of the discourse, which is finally translated into natural language to be presented to the user. After this, the system waits for an answer from the user, indicating if the explanation provided satisfies him/her. If not, the process is repeated.

Then, the objective of this method is *comprehension* of both the *model* and the *reasoning*. Explanations are generated at the *macro* or *micro* level, depending on the user needs, by a *dialog* expressed in *natural language* between the user and the system. Moreover, the *user model* is dynamic, which lets the system update automatically the *level of detail* of the explanation.

3.2.2 Dialog planning

The ideas proposed by Cawsey (1995), related to the generation of interactive explanations, are very similar. The key idea consists not only of planning the text to be presented to the user but also planning the dialog with the user. The main purpose of this method is the *comprehension* of the *model* and the *reasoning*. Explanations are generated at a *macro* or *micro* level, depending on the user needs, by a *dialog* expressed in *natural language* between the user and the system. Moreover, the *user model* is dynamic, which allows the system to update automatically the *level of detail* of the explanation.

There are some expert systems which have implemented those ideas:

EDGE This expert system, developed by Cawsey (1994, 1995) to generate tutorials about electronic circuits, provides the user with the possibility to interrupt the explanation and then to update the explanation. The system consists of a set of planning rules and a planner to generate the dialog. A planning rule has six elements: name, arguments, constraints, preconditions, subgoals and a template for describing the topic. To plan the dialog, each objective is assigned a priority factor. At each moment, the objective with highest priority is selected in order to choose a planning rule to find new objectives. In this way, a hierarchical structure is created and taken as the base of the context in which the explanation is generated. The interruptions made by the user can generate new objectives as well as serve as a way for the system to update the level of knowledge about him/her.

P.rex The same ideas have been implemented in *P.rex* (Fiedler, 2001b), an expert system for explaining the proofs of mathematical theorems. This system allows the user to interact with it at every moment. Also the system dialogs with the user in order to know if he/she has understood the explanation.

Also related to dialog planning and explanation is the theory developed by Grosz & Sidner (1986). They differentiate three components of discourse structure as the basis for explaining interruptions, cue phrases and referring expressions: the structure of the sequence of words or expressions in the discourse, its intention and the state of focus of attention of the participants in the discourse. The many issues related to these topics constitute their current research, especially the aspects involved in collaboration (Grosz & Kraus, 1996) and decision making (Grosz *et al.*, 2004). The main purpose of this theory is the *comprehension* of the *reasoning* in order to explain at a *macro* or *micro* level the processing of utterances in a discourse expressed in *natural language*, taking into account different levels of knowledge of the users about the domain.

3.3 Argumentation

Another area of research related to explanation is argumentation as a tool for generating interactive and collaborative explanations (Moulin *et al.*, 2002). Although both concepts seem very similar (in fact, some authors (Moore & Swartout, 1990) have used argumentation for generating explanations), the main difference between them is that in explanation the topic to be explained is not discussed, whereas in argumentation it is (Vries *et al.*, 2002). However, as explanations tend to support more interactivity between the explainer and the explainee, the use of critics that test their credibility could be more useful (Moulin *et al.*, 2002). Taking into account that argumentative techniques and models have been used in many different application domains (Moulin *et al.*, 2002), in this paper we will focus on the case of logical reasoning in which conclusions are backed up by the reasons for getting them. Nevertheless, a good review about the topic of argumentation is found in Moulin *et al.* (2002), in which a framework which unifies different approaches is also described.

In this section we describe the most relevant works in the area of logical reasoning, depending on the method used to present the argumentation to the user, because all of them have many other properties in common: the objective of the methods is *comprehension* of both the *model* and the *reasoning*, explanations are generated at the *macro* or *micro* level, depending on the user needs, usually presented by a *dialog* expressed in *natural language*. Moreover, the *user model* is dynamic, which allows the system to automatically update the *level of detail* of the explanation.

3.3.1 Textual argumentation

Krause *et al.* (1993/4) developed a qualitative method for carcinogenic risk assessment based on a computational model of argumentation (Krause *et al.*, 1995). The main objective of the system is the generation of a report as informative as possible as well as a 'clear' indication of the reliability of that report and the sources of uncertainty. The idea is to translate the structures of arguments (represented by λ-terms) to text expressed in natural language. Moreover, the user has the ability to request an explanation of any of the phrases in the report and then the system generates new explanations for the selected phrase following the same concept.

3.3.2 Graphical argumentation

It has been shown that graphical representations of the structure and links of an argumentation improve the human ability to solve ill-structured, real world problems, as applied to the representation of *design rationale* underlying collaborative software design decisions (Buckingham-Shum *et al.*, 1997). A design rationale is a way of representing the reasoning associated to a design process. Although there are several graphical notations (*CSCARS*, 2004), Buckingham-Shum *et al.* (1997) have demonstrated empirically that the use of the QOC notation[3] (McLean *et al.*, 1989) for graphical argumentation helps in the design process of *wicked problems* (Rittel & Webber, 1973), but it is quite misleading when the design space is well constrained. Therefore, some techniques are needed to express such notation into a more intelligible document, such as a hypertext report or a digital document, in order to facilitate the understanding of the argumentation.

4 Conclusions

It is widely acknowledged that nowadays explanation constitutes a key issue to be addressed by any expert system in order to be accepted by every kind of user. In fact, explanation can be used for educational purposes in tutorial systems, or to increase the confidence in the results provided by the

[3] *Questions* to represent the key issues of the design, *Options* for answering the Questions, and *Criteria* for assessing one Option over another.

system; even to help the knowledge engineer to build and debug the system. However, it is not very clear what an explanation should include to satisfy the user.

In this paper (Section 2) we have summarized the main properties of explanation (see also Lacave & Díez, 2002). In the light of these, we have analysed explanation methods proposed to date for heuristic expert systems and we have classified those methods into two main groups, depending on their ability to automatically adapt their behaviour to different user requirements. Non-adaptive methods have been presented in Section 3.1 and those that are capable of communicating with the user were given in Section 3.2. Finally, Section 3.3 related explanation to argumentation and described the main work in this area.

4.1 Main ideas

Two main focuses can be observed: the one which considers explanation as a different and independent task from reasoning (Barzilay *et al*., 1998; Wick & Slagle, 1989), which implies that the system needs to represent the explanation knowledge, versus all the rest in which explanations are generated by interacting with the domain knowledge.

However, there are a variety of techniques, ranging from the most rudimentary, like canned text (Bleich, 1972), to the most sophisticated in which explanation constitutes an intelligent dialog with the user (Moore, 1994; Cawsey, 1995). Among these techniques, there are a variety of methods trying to present explanations in natural language. In fact, it seems that the problem of natural language generation is one of the most relevant issues for future research. In this sense, the concept of translating the rules used by a rule-based system has been used not only by MYCIN (Buchanan & Shortliffe, 1984) but also in other systems (Patil *et al.,* 1982; Clancey, 1983; Swartout, 1983; Wallis & Shortliffe, 1984). Nevertheless, the most recent proposals include new techniques of *text planning* that take into account the user model for tailoring the explanation (Marco *et al*., 1995; Carolis *et al*., 1996) and *argumentation* (Moulin *et al*., 2002).

4.2 Comparison with explanation in Bayesian networks

It is puzzling that explanation methods for Bayesian networks (Lacave & Díez, 2002), have not addressed (to date) two important aspects which have already been addressed in the case of heuristic expert systems: the *interaction* between the user and the system, and the possibility of *adaptation*. In the first case, interaction is usually made by windows and menus through a graphical interface and the ability for adaptation is reduced to modify some importance threshold or some level of detail. However, there are some interesting proposals for explanation in Bayesian networks which could be taken into account to improve explanations in any kind of expert system, such as the graphical display of the *chains of reasoning*, the *qualitative representation* of quantitative data or the possibility of analysing both the model and the results provided by the system by different techniques of *sensitivity analysis*. In general, the graphical representation of Bayesian networks provides by itself a powerful tool for explanation, as well as being very intuitive for any kind of user. In fact, we think that the absence of graphical explanations constitutes a big gap in heuristic expert systems.

However, in spite of the limitations of the methods for explanation in Bayesian networks, their use for automatic argument generation in any expert system seems quite promising (Zukerman *et al*., 1999, 2000).

4.3 Future work

This review has summarized a good number of very interesting proposals for generating explanations for expert systems. Nevertheless, there is still much research to be done. In our opinion, the most promising lines are related to the generation of *natural language* explanations as well as the application of *user models*, both for the domain knowledge and for the knowledge about the reasoning method. Moreover, presentation of explanations should be improved by combining different *multimedia resources*, according to the new capabilities provided by recent technological advances and the new ways of interacting with computers (Carroll, 2001), because it has been proved that this constitutes an incentive for non-expert users. In fact, in most explanation methods the interaction with the user is usually done by predefined questions in natural language. However, in real life, this is not the case. In particular, the receiver of the explanation should have the opportunity to express him/herself such questions or indications to be better explained, letting him/her communicate in different ways: textually, or graphically or, even, combining multimedia resources. Therefore, a very promising line for future work is open for the development of new *interaction techniques*.

Acknowledgements

This work has been partially supported by the Spanish CICYT under projects TIC-97-1135-C04 and TIC-2001-2973-C05. We would also like to thank Simon Parsons, Barbara Grosz and Bernard Moulin their help with some references.

References

Barzilay, R, McCullough, D, Rambow, O, de Christofaro, J, Korelsky, T and Lavoie, B, 1998, A new approach to expert system explanations. In Hovy, E (ed.), *Proceedings of the Ninth International Workshop on Natural Language Generation*. New Brunswick, NJ: Association for Computational Linguistics, pp. 78–87.

Bennet, S and Scott, A, 1984, Specialized explanations for dosage selection. In Buchanan, B. and Shortliffe, E., *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley, pp. 363–370.

Bleich, H, 1972, Computer-based consultation: Electrolyte and acid-base disorders. *American Journal of Medicine* **53**, 285–291.

Buchanan, B and Shortliffe, E, (eds.) 1984, *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley.

Buckingham-Shum, S, MacLean, A, Bellotti, V and Hammond, N, 1997, Graphical argumentation and design cognition. *Human Computer Interaction* **12**, 267–300.

Carenini, G, Mittal, V and Moore, J, 1994, Generating patient-specific interactive natural language explanations. In *Proceedings of 18th Symposium on Computer Applications in Medical Care (SCAMC-94)*. New York: McGraw-Hill.

Carolis, BD, Rosis, FD, Grasso, F, Rossiello, A, Berry, D and Gillie, T, 1996, Generating recipient-centered explanations about drug prescription. *Artificial Intelligence in Medicine* **8**, 123–145.

Carroll, J, 2001, *Human–Computer Interaction in the New Millennium*. New York: Addison-Wesley Professional.

Cawsey, A, 1993, User-modelling in interactive explanations. *Journal of User Modelling and User Adapted Interaction* **3**(3), 221–247.

Cawsey, A, 1994, Developing an explanation component for a knowledge-based system: Discussion. *Expert Systems with Applications* (special issue on explanation) **8**(4), 527–531.

Cawsey, A, 1995, *Explanation and Interaction*. Cambridge, MA: MIT Press.

Chandrasekaran, B, Tanner, M and Josephson, J, 1989, Explaining control strategies in problem solving. *IEEE Expert* **4**, 9–24.

Clancey, W, 1983, The epistemology of rule-based expert systems. A framework for explanation. *Artificial Intelligence* **20**, 215–251.

Clancey, W, 1993, Notes on 'Heuristic classification'. *Artificial Intelligence* **59**, 191–196.

CSCARS 2004, `http://kmi.open.ac.uk/people/sbs/csca/`

Doyle, D, Tsymbal, S and Cunningham, P, 2003, A review of explanation and explanation in case-based reasoning. Technical Report TCD-CS-2003–41, Department of Computer Science, Trinity College, Dublin.

Druzdzel, M, 1993, Probabilistic reasoning in decision support systems: From computation to common sense. PhD Thesis, Department of Engineering and Public Policy, Carnegie Mellon University.

Fiedler, A, 2001a, Dialog-driven adaptation of explanations of proofs. In *Proceedings of the Seventeenth International Joint Conference of Artificial Intelligence, Seattle, Washington*, pp. 1295–1300.

Fiedler, A, 2001b, P.rex: An interactive proof explainer. In *Proceedings of the First International Conference on Automated Reasoning, Siena, Italy*.
Grosz, B and Kraus, S, 1996, Collaborative plans for complex group action. *Artificial Intelligence* **86**(2), 269–357.
Grosz, B, Kraus, S, Talman, S, Stossel, B and Havlin, M, 2004, The influence of social dependencies on decision-making: Initial investigations with a new game. In *Proceedings of Autonomous Agents and Multi-Agent Systems*, pp. 782–789.
Grosz, B and Sidner, C, 1986, Attention, intentions and the structure of discourse. *Computational Linguistics* **12**(3), 175–204.
Henrion, M, 1989, Some practical issues in constructing belief networks. In Henrion, M, Shachter, R, Kanal, L and Lemmer, J, (eds.), *Uncertainty in Artificial Intelligence*, vol. 5. Elsevier: North-Holland, pp. 161–173.
Henrion, M and Druzdzel, M, 1990, Qualitative propagation and scenario-based approaches to explanation of probabilistic reasoning. In *Proceedings of the 6th Conference on Uncertainty in Artificial Intelligence*, pp. 17–32.
Horvitz, E, Heckerman, D, Nathwani, B and Fagan, L, 1986, The use of a heuristic problem-solving hierarchy to facilite the explanation of hypothesis-directed reasoning. In *MEDINFO 86*. New York: Elsevier, pp. 27–31.
Kahneman, D, Slovic, P and Tversky, A, (eds.) 1982, *Judgement Under Uncertainty: Heuristics and Biases*. Cambridge, UK: Cambridge University Press.
Krause, P, Ambler, S and Fox, J, 1995, A logic of argumentation for uncertain reasoning. *Computational Intelligence* **11**(1), 113–131.
Krause, P, Fox, J and Judson, P, 1993/4, An argumentation based approach to risk assessment. *IMA Journal of Mathematics Applied in Business and Industry* **5**, 249–263.
Lacave, C and Díez, F, 2002, A review of explanation methods for Bayesian networks. *The Knowledge Engineering Review* **17**(2), 107–127.
Lacave, C and Díez, FJ, 2003, Knowledge acquisition in Prostanet, a Bayesian network for diagnosing prostate cancer. In *Knowledge-Based Intelligent Information and Engineering Systems (Lecture Notes in Artificial Intelligence, 2774)*. Oxford, UK: Springer, pp. 1345–1350.
Langlotz, C, Shortliffe, E and Fagan, L, 1988, A methodology for generating computer-based explanations of decision-theoretic advice. *Medical Decision Making* **8**(4), 290–303.
Marco, CD, Hirst, G, Wanner, L and Wilkinson, J, 1995, HealthDoc: Customizing patient information and health education by medical condition and personal characteristics. In *First International Workshop on Artificial Intelligence in Patient Education, Glasgow, UK*.
McGuinness, D and Pinheiro da Silva, P, 2003, Infrastructure for web explanations. In Fensel, D, Sycara, K & Mylopoulos, J, (eds.), *Proceedings of the 2nd International Semantic Web Conference (ISWC2003) (Lecture Notes in Computer Science, 2870)*. Sanibel, FL: Springer, pp. 113–119.
McLean, A, Young, R and Moran, T, 1989, Design rationale: the argument behind the artifact. In *Proceedings of the CHI'89 Conference on Human Factors in Computing Systems, New York*, pp. 247–252.
Moore, J, 1994, *Participating in Explanatory Dialogues: Interpreting and Responding to Questions in Context*. Cambridge, MA: MIT Press.
Moore, J and Swartout, W, 1990, Pointing: A way toward explanation dialogue. In *Proceedings of the 8th National Conference on Artificial Intelligence, Boston, MA*, pp. 457–464.
Moulin, B, Irandoust, H, Bélanger, M and Desbordes, G, 2002, Explanation and argumentation capabilities: Towards the creation of more persuasive agents. *Artificial Intelligence Review* **17**(3), 169–222.
Patil, R, Szolovits, P and Schwartz, W, 1982, Modeling knowledge of the patient in acid–base and electrolyte disorders. *Artificial Intelligence in Medicine* 345–348.
Pearl, J, 1988, *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. San Mateo, CA: Morgan Kaufmann Publishers.
Pearl, J, 2001, *Causality. Models, Reasoning and Inference*. Cambridge, UK: Cambridge University Press.
Pennington, N and Hastie, R, 1988, Explanation-based decision making: effects of memory structure on judgment. *Journal of Experimental Psychology: Learning, Memory and Cognition* **14**(3), 521–533.
Reggia, J and Perricone, B, 1985, Answer justification in medical decision support systems based on Bayesian classification. *Computers in Biology and Medicine* **15**(4), 161–167.
Rittel, H and Webber, M, 1973, Dilemmas in a general theory of planning. *Policy Sciences* **4**, 155–169.
Scott, A, Clancey, W, Davis, R and Shortliffe, E, 1984, Methods for generating explanations. In Buchanan, B and Shorliffe, E, (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley, pp. 338–362.
Sember, P and Zukerman, I, 1990, Strategies for generating micro explanations for Bayesian belief networks. In Bonissone, P, Henrion, M, Kanal, L and Lemmer, JF, (eds.), *Uncertainty in Artificial Intelligence* **5**. Amsterdam: Elsevier, pp. 295–302.

Shortliffe, EH, Axline, S, Buchanan, BG, Merigan, T and Cohen, S, 1973, An artificial intelligence program to advise physicians regarding antimicrobial therapy. *Computers and Biomedical Research* **6**, 544–560.

Silverman, H, 1975, A digitalis therapy advisor. Technical Report, MIT AI Lab, January.

Strat, T, 1987, The generation of explanations within evidential reasoning systems. In *Proceedings of the 10th International Joint Conference on Artificial Intelligence, Milan, Italy*, pp. 1097–1104.

Suermondt, H, 1992, Explanation in Bayesian works. PhD Thesis, Department of Computer Science, Stanford University, Stanford, CA, STAN-CS-92–1417.

Swartout, W, 1983, XPLAIN: A system for creating and explaining expert consulting programs. *Artificial Intelligence* **21**, 285–325.

Teach, RL and Shortliffe, EH, 1984, An analysis of physician's attitudes. In Buchanan, B and Shorliffe, E, (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley, pp. 635–652.

Vries, ED, Lund, K and Baker, M, 2002, Computer-mediated epistemic dialogue: explanation and argumentation as vehicle for understanding scientific notions. *The Journal of Learning Sciences* **11**(1), 63–103.

Wallis, J and Shortliffe, E, 1984, Customized explanations using causal knowledge. In Buchanan, B and Shorliffe, E, (eds.), *Rule-Based Expert Systems: The MYCIN Experiments of the Stanford Heuristic Programming Project*. Reading, MA: Addison-Wesley, pp. 371–388.

Wick, M, 1989, The 1988 Workshop on Explanation. *Artificial Intelligence Magazine* **10**(3), 22–26.

Wick, M and Slagle, J, 1989, An explanation facility for today's expert systems. *IEEE Expert* **4**, 24–36.

Wick, M and Thompson, W, 1992, Reconstructive expert system explanation. *Artificial Intelligence* **54**, 33–70.

Wooley, B, 2000, Explanation component of software systems. *Crossroads. The ACM's First Electronic Publication*, `http://www.acm.org/crossroads/xrds5-1/explain.html`.

Ye, R and Johnson, PE, 1995, The impact of explanation facilities on user acceptance of expert system's advice. *MIS Quarterly* **19**(2), 157–172.

Zukerman, I, McConachy, R and Korb, R, 2000, Using argumentation strategies in automated argument generation. In *Proceedings of the First International Natural Language Generation Conference, Mitzpe Ramon, Israel*, pp. 55–62.

Zukerman, I, McConachy, R, Korb, R and Pickett, D, 1999, Exploratory interaction with a Bayesian argumentation system. In MK Publishers (ed.), *Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Stockholm, Sweden*, pp. 1294–1299.